Improving Energy Efficiency in Femtocell Networks: A Hierarchical Reinforcement Learning Framework

Xianfu Chen*, Honggang Zhang†‡, Tao Chen*, and Mika Lasanen*

*VTT Technical Research Centre of Finland, P. O. Box 1100, FI-90571 Oulu, Finland

†York-Zhejiang Lab for Cognitive Radio and Green Communications

‡Department of ISEE, Zhejiang University, Zheda Road 38, Hangzhou 310027, China

Email: xianfu.chen@vtt.fi, honggangzhang@zju.edu.cn, tao.chen@vtt.fi, mika.lasanen@vtt.fi

Abstract—**This paper investigates energy efficiency for two-tier femtocell networks through combining game theory and stochastic learning. With the Stackelberg game formulation, a hierarchical reinforcement learning framework is applied to study the joint average utility maximization of macrocells and femtocells subject to the minimum signal-to-interference-plus-noise-ratio requirements. The macrocells behave as the leaders and the femtocells are followers during the learning procedure. At each time step, the leaders commit to dynamic strategies based on the best responses of the followers, while the followers compete against each other with no further information but the leaders' strategy information. In this paper, we propose two learning algorithms to schedule each cell's stochastic power levels, leading by the macrocells. Numerical experiments are presented to validate the proposed studies and show that the two learning algorithms substantially improve the energy efficiency of the femtocell networks.**

I. INTRODUCTION

The rapidly rising energy consumption and greater awareness of environmental protection have created an urgent need for energy-efficient wireless communication techniques. In wireless cellular networks, the radio access part is the main source of energy consumption, accounting for up to more than 70% [1]. Therefore, it's crucial to increase the energy efficiency in wireless access networks to meet the challenges raised by the exponential growth in mobile services and energy consumption.

Recently, femtocells have been considered as a promising technique for improving the indoor experience of the mobile users. Due to the short transmit-receive distance property, femtocell technique can greatly lower transmit power, prolong handset battery life, and achieve improved reception [2]. Although femtocells bring significant benefits to both mobile operators and costumers, to make the deployment feasible, there still exist many technical challenges, e.g., cross-tier/co-tier interference management. In [3], the authors proposed a distributed utility-based signal-to-interference-plus-noise ratio (SINR) adaptation algorithm to alleviate the cross-tier interference at the macrocell from co-channel femtocells. In [4], a Stackelberg game was formulated to study the resource allocation in two-tier femtocell networks, where the macro base station (MBS) protects itself by pricing the interference from femtocell users (FUs).

This paper discusses the energy efficiency in femtocell networks. In [5], the problem of energy-efficient spectrum sharing and power allocation in cognitive radio femtocells was studied, where a three-stage Stackelberg game model was formulated to improve the energy efficiency. In [6], the authors proposed a novel energy saving procedure for the femtocell base station (FBS) to decide when to switch on/off. Moreover, we focus mainly on discussing the co-channel operation of femtocells with closed access. The interference in this scenario can not be handled by means of centralized network scheduling, because the number and locations of femtocells are unknown. For such considered networking environment, the femtocells are most likely to be autonomous. Existing research works on reinforcement learning (RL) [7] have been carried out in femtocelss networks. In [8], a realtime multi-agent RL algorithm that optimizes the network performance by managing the interference in femtocell networks was proposed. In [9], a distributed learning scheme based on Q-learning was developed to manage the femto-to-macrocell cross-tier interference in femtocell networks.

In this paper, we model the energy efficiency aspect of power allocation problem in femtocell networks as a Stackelberg learning game, or leader-follower learning process, with the following characteristics: 1) the macrocells are considered to be the leaders, whereas the femtocells are considered to be the followers; 2) the leaders behave by knowing the response of the femtocells to their own strategy decisions; 3) given the leaders' decisions, the followers compete with each other. Learning is accomplished by directly interacting with the surrounding environment and properly adjusting the strategies according to the realizations of achieved performance. The solution of such a learning game is the Stackelberg equilibrium (SE). If no hierarchy exists during the learning procedure, the Stackelberg learning game reduces to the non-cooperative learning game, which is the scenario discussed in [10].

The rest of this paper is organized as follows. The next section presents the energy efficiency problem in femtocell networks. Section III defines a Stackelberg game theoretic
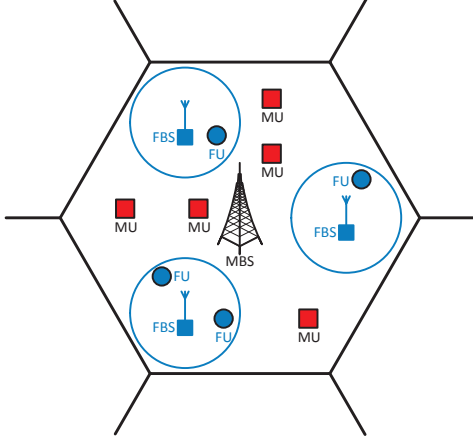
Fig. 1. Macrocell is underlaid with femtocells located in the coverage of the macrocell (MBS: macro base station; MU: macrocell user; FBS: femto base station; FU: femtocell user.).

solution for the users' hierarchical behaviors. In Section IV, a Stackelberg learning framework is proposed, and two RL based algorithms are further derived. The numerical results are included in Section V, verifying the validity and efficiency of the proposed algorithms. Finally, we present in Section VI a conclusion of this paper.

II. PROBLEM FORMULATION

The network scenario considered in this paper is shown in Fig. 1, where there exist multiple femtocells and macrocells. Each macrocell consisting of a MBS and multiple macro users (MUs), is underlaid with several co-channel femtocells. In each femtocell, there is one FBS providing service to femtocell users (FUs). Here we assume the closed policy, since customers may prefer that policy because of privacy concerns and limited backhaul bandwidth. Assume same distribution of femtocells in each macrocell, we concentrate on the case of one macrocell for simplification without loss of generality. Suppose N femtocells $B_i(i \geq 1)$ operate within the coverage of the macrocell B_0. Users of the same macrocell/femtocell adopt time division multiple access (TDMA) for data transmission, thus causing no interference for each other. In the following, this paper mainly addresses the uplink transmissions in the cell over a common spectrum band.

Let $i \in \mathcal{N}$ denote the scheduled user connected to its BS B_i, where $\mathcal{N} = \{0, 1, \ldots, N\}$ refers to the user index set. Designate the power level of user i as $p_i \in P_i \stackrel{\text{def}}{=} [p_i^{\min}, p_i^{\max}]$, the SINR γ_i of user i at B_i is given by

$$\gamma_i \left(p_i, \mathbf{p}_{-i} \right) = \frac{h_{i,i} p_i}{\sum_{j \in \mathcal{N} \backslash \{i\}} h_{j,i} p_j + \sigma^2}.$$

Here $-i$ denote all the users in the set $\mathcal{N}$ except user i, σ^2 is the variance of background Additive White Gaussian Noise (AWGN), and $\{h_{j,i}\}$ is the set of channel gains from user j to B_i. Each user aims to improve the energy efficiency, which

can be defined as

$$\xi_i \left(p_i, \mathbf{p}_{-i} \right) = \frac{W \log_2 \left(1 + \gamma_i \left(p_i, \mathbf{p}_{-i} \right) \right)}{p_i^a + p_i},$$

where W is the spectrum bandwidth, and p_i^a denotes the additional circuit power consumption of user i during transmissions and is independent from the transmission power.

Given the minimum SINR requirement γ_i^* chosen to ensure the adequate QoS, we can express the utility function of the user i formally as

$$u_i \left(p_i, \mathbf{p}_{-i} \right) = \begin{cases} \xi_i \left(p_i, \mathbf{p}_{-i} \right), & \text{if } \gamma_i \left(p_i, \mathbf{p}_{-i} \right) \geq \gamma_i^*; \\ 0, & \text{otherwise.} \end{cases}$$

Whenever the macrocell's γ_0^* becomes infeasible, MU transmits with maximum power level. We propose that the MBS requests the FUs to reduce their minimum SINR targets. Each user $i \in \mathcal{N}$ tries to find the optimal power level p_i that maximizes its utility,

$$\max_{p_i \in P_i} u_i \left(p_i, \mathbf{p}_{-i} \right). \tag{1}$$

III. STACKELBERG GAME THEORETIC SOLUTION

Within the Stakelberg game framework, the MU is modeled as the leader, while the FUs as the followers competing with each other over the spectrum. In order to investigate the existence of an SE, we first define p_i^* to be the best response to $\mathbf{p}_{-i}$ if

$$u_i \left(p_i^*, \mathbf{p}_{-i} \right) \geq u_i \left(p_i, \mathbf{p}_{-i} \right), \forall p_i \in P_i.$$

User i's best response to $\mathbf{p}_{-i}$ is denoted as $BR_i(\mathbf{p}_{-i})$. Let $NE(p_0)$ be the Nash equilibrium (NE) strategy of the FUs if the MU chooses to play p_0, i.e.

$$NE(p_0) = \mathbf{p}_{-0}, \text{ if } p_i = BR_i(\mathbf{p}_{-i}), \forall i \in \mathcal{N} \backslash \{0\}.$$

The strategy profile $(p_0^*, NE(p_0^*))$ is an SE if and only if

$$u_0 \left(p_0^*, NE(p_0^*) \right) \geq u_0 \left(p_0, NE(p_0) \right), \forall p_0 \in P_0.$$

The following theorem establishes the existence of the SE.

Theorem 1. The SE always exists in our defined Stackelberg game with the MU leading and the FUs following.

Proof: For $\forall p_0 \in P_0$, there is at least one NE in the non-cooperative game $G = \langle p_0, \mathcal{N} \backslash \{0\}, \{P_i\}, \{u_i\} \rangle$, since for $\forall i \in \mathcal{N} \backslash \{0\}$

1) the strategy profile $\mathcal{P}_i$ is a non-empty, convex, and compact subset of some Euclidean space $\mathfrak{R}^n$;
2) u_i is continuous in $\mathbf{p}_{-i}$ and quasi-concave in p_i.

Therefore, there exists $p_0^* \in P_0$, such that

$$u_0 \left(p_0^*, NE \left(p_0^* \right) \right) = \sup_{p_0 \in P_0} u_0 \left(p_0, NE(p_0) \right).$$

We can conclude that the SE always exists in the defined Stackelberg game. ∎

Note that there is only one leader in the Stackelberg game. The MU regards itself as the leader and performs the Stackelberg strategy, and the FUs will act their best responses until reach the equilibrium.

IV. STACKELBERG LEARNING FRAMEWORK

In the Stackelberg learning game, each user in the network behaves as an intelligent agent, whose objective is to maximize its expected utility. The game is played repeatedly to achieve the optimal strategies. The Stackelberg learning framework has two levels of hierarchy: 1) the MU aims to maximize its expected utility by knowing the responses of all FUs for each possible play; 2) given the strategy of the MU, the FUs play a non-cooperative game among each other. In this section, we focus our emphasis on how to reach the optimal communication configuration through RL approach [7].

A strategy for user $i \in \mathcal{N}$ at time step t is defined to be a probability vector $\mathbf{y}_i^t = (y_{i,1}^t, \ldots, y_{i,m_i}^t) \in \mathcal{Y}_i$, where $y_{i,j}^t$ means the probability with which the user i chooses action (transmission power) $p_{i,j} \in \mathcal{P}_i$, and $\mathcal{Y}_i$ is the strategy set available to user i. Since each user can only choose a power level from a finite discrete set, $\mathcal{P}_i$ is assumed to be a finite set with dimension m_i. Then the expected utility for user i at time slot t can be expressed as follows

$$U_i \left(\mathbf{y}_i^t, \mathbf{y}_{-i}^t \right) = \mathrm{E} \left[u_i | \text{user } j \text{ plays strategy } \mathbf{y}_j^t, j \in \mathcal{N} \right]$$
$$= \sum_{\mathbf{p}^t \in \mathcal{P}} u_i \left(\mathbf{p}^t \right) \prod_{s \in \mathcal{N}} y_{s,j_s}^t,$$

where $\mathbf{p}^t = (p_{0,j_0}^t, \ldots, p_{N,j_N}^t)$ is the vector of actions chosen by all users at time t, and $\mathcal{P} = \times_{i \in \mathcal{N}} \mathcal{P}_i$ is the set of all action vectors.

For any stationary strategy[1] of the MU $\mathbf{y}_0 \in \mathcal{Y}_0$, the best-response strategies of all FUs define an NE strategy $NE(\mathbf{y}_0)$, i.e. $NE(\mathbf{y}_0) = \mathbf{y}_{-0}^*$, if

$$\mathbf{y}_i^* = \arg \max_{\mathbf{y}_i \in \mathcal{Y}_i} U_i \left(\mathbf{y}_i, \mathbf{y}_{-i} \right), \forall i \in \mathcal{N} \backslash \{0\}.$$

The MU's optimal strategy is then

$$\mathbf{y}_0^* = \arg \max_{\mathbf{y}_0 \in \mathcal{Y}_0} U_0 \left(\mathbf{y}_0, NE(\mathbf{y}_0) \right).$$

Together $(\mathbf{y}_0^*, NE(\mathbf{y}_0^*))$ constitute a stationary strategy of SE for the Stackelberg learning formulation.

Theorem 2. For the considered Stackelberg learning game, there exist a MU's stationary strategy and a FUs' NE strategy that form an SE.

Inspired by [11], we can prove Theorem 2 as follows.

Proof: If the MU follows a stationary strategy $\mathbf{y}_0$, then the Stackelberg learning game is simplified to be a N-player stochastic learning game. It has been shown in [12] that every finite strategic-form game has a mixed strategy equilibrium, that is, there always exists a $NE(\mathbf{y}_0)$ in our formulation of the discrete power allocation game given the MU's strategy $\mathbf{y}_0$. The rest of the proof follows directly from the definition of SE, and is thus omitted for brevity. ∎

[1]A strategy is said to be stationary, where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m_i})$ is not changing with time during the stochastic learning process.

A. Reinforcement Learning based Algorithm-I (RLA-I)

During the Stackelberg learning process, the MU behaves as the leader and knows the strategy information of all FUs. Users with RL ability learn to maximize its individual expected utility through repeated interactions with the networking environment. Among many different implementation of above adaptation mechanism, here we consider the so-called Q-learning [13], where the users' strategies are parameterized through Q-functions that characterize relative expected utility of a particular power level. More specifically, let $Q_i^t (p_{i,j_i})$ denote the Q-value of user i's corresponding power level p_{i,j_i} at time t. Then, after selecting power level p_{i,j_i} at time $t+1$, the Q-value is updated according to

$$Q_i^{t+1} \left(p_{i,j_i} \right) = Q_i^t \left(p_{i,j_i} \right) + \alpha \left(U_i \left(p_{i,j_i}, \mathbf{y}_{-i}^{t+1} \right) - Q_i^t \left(p_{i,j_i} \right) \right), \tag{2}$$

where $\alpha \in [0,1)$ is the learning rate, and

$$U_i \left(p_{i,j_i}, \mathbf{y}_{-i}^{t+1} \right) = \sum_{\mathbf{p}_{-i}^{t+1} \in \mathcal{P}_{-i}} u_i \left(p_{i,j_i}, \mathbf{p}_{-i}^{t+1} \right) \prod_{s \in \mathcal{N} \backslash i} y_{s,j_s}^{t+1}. \tag{3}$$

Each FU $i \in \mathcal{N} \backslash \{0\}$ doesn't know other competing FUs' private strategies $\mathbf{y}_{-(0,i)}^{t+1}$, and the only information it has is the MU's transmission parameters. But the FU i is able to compute the attainable utility $u_i(p_{i,j_i}, \mathbf{p}_{-i}^{t+1})$ with the feedback information from its receiver. Therefore, FU i can estimate $U_i(p_{i,j_i}, \mathbf{y}_{-i}^{t+1})$ using

$$\widetilde{U}_i^{t+1} \left(p_{i,j_i} \right) = \sum_{p_{0,j_0}^{t+1} \in \mathcal{P}_0} y_{0,j_0}^{t+1} \widehat{U}_i^{t+1} \left(p_{i,j_i}, p_{0,j_0}^{t+1} \right), \tag{4}$$

where $\widehat{U}_i^{t+1} \left(p_{i,j_i}, p_{0,j_0}^{t+1} \right)$ is given in Eq. (5). $n_{p_{i,j_i}, p_{0,j_0}}^t$ is the number of times when MU selects power level p_{0,j_0} and FU i selects power level p_{i,j_i} until time slot t. The updating rule in Eq. (2) can then be rewritten as

$$Q_i^{t+1} \left(p_{i,j_i} \right) =$$
$$\begin{cases} Q_i^t \left(p_{i,j_i} \right) + \alpha \left(U_i \left(p_{i,j_i}, \mathbf{y}_{-i}^{t+1} \right) - Q_i^t \left(p_{i,j_i} \right) \right), & \text{if } i = 0; \\ Q_i^t \left(p_{i,j_i} \right) + \alpha \left(\widetilde{U}_i^{t+1} \left(p_{i,j_i} \right) - Q_i^t \left(p_{i,j_i} \right) \right), & \text{otherwise.} \end{cases}$$
$$\tag{6}$$

Then, we need to specify how the user selects power levels. Greedy selection, when the action with the highest Q-value is selected, might generally lead to globally suboptimal solution. Thus, we need to incorporate some way of exploring less-optimal strategies. Here, we focus on Boltamann action selection mechanism, that is, the probability of choosing power level p_{i,j_i} at time t is given by

$$y_{i,j_i}^{t+1} = \frac{\exp \left(Q_i^t \left(p_{i,j_i} \right) / \tau_i \right)}{\sum_{l=1}^{m_i} \exp \left(Q_i^t (p_{i,l}) / \tau_i \right)}, \tag{7}$$

where the temperature τ_i controls the exploration/exploitation tradeoff [7].

Now we present the reinforcement learning based algorithm-I for the Stackelberg learning game.

RLA-I

$$\widehat{U}_i^{t+1}\left(p_{i,j_i}, p_{0,j_0}^{t+1}\right) = \begin{cases} \dfrac{u_i\left(p_{i,j_i}, p_{0,j_0}, \mathbf{p}_{-(i,0)}^{t+1}\right) - \widehat{U}_i^t\left(p_{i,j_i}, p_{0,j_0}\right)}{n_{p_{i,j_i},p_{0,j_0}}^t + 1} + \widehat{U}_i^t\left(p_{i,j_i}, p_{0,j_0}\right), & \text{if } p_{0,j_0}^{t+1} = p_{0,j_0}; \\ \widehat{U}_i^t\left(p_{i,j_i}, p_{0,j_0}^{t+1}\right), & \text{otherwise.} \end{cases} \quad (5)$$

Initialization:

$t = 0$, initialize Q-values $Q_i^t(p_{i,j_i})$, for $\forall i \in \mathcal{N}$ and $\forall p_{i,j_i} \in \mathcal{P}_i$.

Learning:

1) The MU chooses action p_{0,j_0} according to y_0^t and broadcasts its strategy information to all the FUs in the network.

2) Each FU $i \in \mathcal{N}\backslash\{0\}$ selects action p_{i,j_i} according to y_i^t and sends its relevant strategy information to the MBS.

3) Users measure their SINR γ_i with the feedback information of the intended receiver. If $\gamma_i \geq \gamma_i^*$, then $\xi_i(\mathbf{p})$ can be achieved; otherwise, the receiver can not receive correctly, thus obtains zero.

4) The MU updates $U_i\left(p_{i,j_i}, \mathbf{y}_{-i}\right)$ according to Eq. (3); and the FUs update $\widetilde{U}_i^t(p_{i,j_i})$ basing on Eq. (4) and Eq. (5).

5) Update Q-values $Q_i^t(p_{i,j_i})$ according to Eq. (6).

6) Update the strategies y_{i,j_i}^{t+1} according to Eq. (7).

7) $t = t + 1$.

End Learning

The learning algorithm results in a stochastic process of choosing power level, so we need to investigate the long-term behavior of the learning procedure. Along with the discussion in [14], we obtain the following differential equation describing the evolution of the Q-values:

$$\frac{\mathrm{d}Q_i^t(p_{i,j_i})}{\mathrm{d}t} = \begin{cases} \alpha\left(U_i\left(p_{i,j_i}, \mathbf{y}_{-i}^t\right) - Q_i^t(p_{i,j_i})\right), & \text{if } i = 0; \\ \alpha\left(\widetilde{U}_i^t(p_{i,j_i}) - Q_i^t(p_{i,j_i})\right), & \text{otherwise.} \end{cases} \quad (8)$$

Next, we would like to express the dynamics in terms of strategies rather than the Q-values. Toward this end, we differentiate Eq. (7) with respect to t and use Eq. (8). The set of equations expressed by Eq. (9) can be arrived.

The first term in braces of Eq. (9) asserts that the probability of choosing power level p_{i,j_i} increases with a rate proportional to the overall efficiency of that strategy, while the second term describes the BS's tendency to randomize over possible power levels. Let $Y^t = (\mathbf{y}_0^t, \ldots, \mathbf{y}_N^t)$ the strategy profile of all users at time t. In the following analysis, we resort to an ordinary differential equation (ODE) whose solution approximates the convergence of Y^t [15]. Then, as $\alpha^t \to 0$, we can represent the right-hand side of Eq. (9) by a function $f(Y^t)$. With any initial $Y^0 = Y_0$, Y^t will converges weakly, as $\alpha \to 0$, to $Y^* \stackrel{\text{def}}{=} (\mathbf{y}_0^*, NE(\mathbf{y}_0^*))$ which is the solution of

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = f(Y), Y^0 = Y_0. \quad (10)$$

Theorem 3. *RLA-I* will always discover an SE strategy.

Proof: We prove this by contradiction. Suppose that the process generated by Eq. (7) converges to a non-Stackelberg equilibrium. We know that the equilibrium solution of Eq. (10) that characterize the long term behavior of *RLA-I* are by definition stationary points. This implies that *RLA-I* will only converge to stationary points. This means that stationary points that are not SEs are stable, which contradicting Theorem 2. ∎

B. Reinforcement Learning based Algorithm-II (RLA-II)

In this subsection, to promote cooperation among the competing FUs, we further design a simple and intuitive rule each FU has that links the other FUs' strategies to its own current transmission strategy. Such a rule reflects an awareness that there are strategic interactions during the learning procedure. FUs with such beliefs do not correctly perceive how the future strategies of their competitors depend on the past. Thus, we propose a conjecture model concerning the way in which the FUs react to each other [16].

Each FU $i \in \mathcal{N}\backslash\{0\}$ thinks any change in its current transmission strategy will induce other FUs to make well-defined changes in the corresponding time slot. Specifically, we need to express FU i's expected contention measure $b_i^{t+1}\left(\mathbf{p}_{-(0,i)}^{t+1}\right) = \prod_{s\in\mathcal{N}\backslash\{0,i\}} y_{s,j_s}^{t+1}$ in Eq. (3) through $\tilde{b}_i^{t+1}\left(\mathbf{p}_{-(0,i)}^{t+1}\right)$, that is

$$\tilde{b}_i^{t+1}\left(\mathbf{p}_{-(0,i)}^{t+1}\right) = b_i^t\left(\mathbf{p}_{-(0,i)}^{t+1}\right) - \delta_i\left(y_{i,j_i}^{t+1} - y_{i,j_i}^t\right), \quad (11)$$

where $\delta_i > 0$ is the belief factor. In other words, every FU i believes that a change of $y_{i,j_i}^{t+1} - y_{i,j_i}^t$ in its strategy at time $t+1$ will induce a change of $\delta_i\left(y_{i,j_i}^{t+1} - y_{i,j_i}^t\right)$ in the expected contention measure exactly related to the transmission strategies of other FUs. It's worth mentioning here that although FU i may be aware that other FUs are subject to many influences on their strategies, when making its own decision, it is only concerned with other FUs' reactions to itself. That means FU i does not take into account whether or not FU j ($j \in \mathcal{N}\backslash\{0,i\}$) might react to changes in transmission strategy made by FU l ($l \in \mathcal{N}\backslash\{0,i,j\}$).

Among different possibilities of capturing the expected contention measure $\tilde{b}_i^{t+1}\left(\mathbf{p}_{-(0,i)}^{t+1}\right)$, the linear model represented in Eq. (11) is the simplest form based on which one FU can model the impact of its changes in transmission strategy to other competing FUs. The conjecture model deployed by the FUs are based on the concept of reciprocity, which refers to interaction mechanisms in which the FUs repeatedly interact when choosing the power level. If they realize that their probabilities of interacting with each other in the future is high, they will consider their influence on the strategies of other FUs, which is captured in the conjecture model by the positive

$$\frac{\mathrm{d}y_{i,j_i}^t}{\mathrm{d}t} = \begin{cases} \dfrac{\alpha}{\tau_0} y_{0,j_0}^t \left\{ \left[U_0\left(p_{0,j_0}, \mathbf{y}_{-0}^{t-1}\right) - \sum\limits_{l=1}^{m_0} y_{0,l}^t U_0\left(p_{0,l}, \mathbf{y}_{-0}^{t-1}\right) \right] - \tau_0 \sum\limits_{l=1}^{m_0} y_{0,l}^t \ln\left(y_{0,j_0}^t / y_{0,l}^t\right) \right\}, & \text{if } i = 0; \\[4mm] \dfrac{\alpha}{\tau_i} y_{i,j_i}^t \left\{ \left[\widetilde{U}_i^{t-1}\left(p_{i,j_i}\right) - \sum\limits_{l=1}^{m_i} y_{i,l}^t \widetilde{U}_i^{t-1}\left(p_{i,l}\right) \right] - \tau_i \sum\limits_{l=1}^{m_i} y_{i,l}^t \ln\left(y_{i,j_i}^t / y_{i,l}^t\right) \right\}, & \text{otherwise.} \end{cases} \tag{9}$$

parameter δ_i. Otherwise, they will act myopically, which is the same learning process as in previous Section IV-A.

Following the previous analysis, Eq. (4) thus becomes

$$\widetilde{U}_i^{t+1}\left(p_{i,j_i}\right) = \sum_{\mathbf{p}_{-i}^{t+1} \in \mathcal{P}_{-i}} y_{0,j_0}^{t+1} u_i\left(p_{i,j_i}, \mathbf{p}_{-i}^{t+1}\right) \tilde{b}_i^{t+1}\left(\mathbf{p}_{-(0,i)}^{t+1}\right), \tag{12}$$

Therefore, we propose *RLA-II* by replacing $\widetilde{U}_i^{t+1}\left(p_{i,j_i}\right)$ in Eq. (6) with Eq. (12) to discover the SE strategy. *RLA-II* is similar to *RLA-I* except that the FUs update their Q-values based on Eq. (4) in *RLA-I*.

V. NUMERICAL RESULTS

We provide insight into the performance comparison of the both learning algorithms through simulation experiments. We consider multiple FUs coexisting with one MU over a spectrum with bandwidth of 1MHz. The minimum SINR targets of MU and FUs are set to be 3dB and 5dB, respectively. The noise of the measurement is according to a zero-mean Gaussian noise with the power of $\sigma^2 = -110$dBm. For simplicity, we suppose that the belief factors δ_i are all equal to 2, $\forall i \in \mathcal{N} \backslash \{0\}$, i.e., the FUs have the same conjecture ability, and the additional circuit power consumption is 10dBm for all users. The femtocells are assumed to be uniformly distributed within a circle centered at the MBS with radius of 500m. The coverage radius of femtocell is 20m. The channel gains are generated by a log-normal shadowing path loss model, $h_{i,j} = d_{i,j}^{-n}$, where $d_{i,j}$ is the distance between user i and BS j, and n is the path loss exponent. In simulation, n is assumed to be 4. The action set of transmission power levels for all users is $\{20, 25, 30\}$dBm.

Fig. 2, Fig. 3 and Fig. 4 show the learning process of expected utilities for each user in the network. The results are compared with

1) The fully cooperative power allocation game with complete information case: each user knows all the utility functions and transmission power levels in the network, and then the optimal utilities in the power allocation process can be calculated by each user according to Eq. (1). This scenario is equivalent to the classic power control game in femtocell networks without the pricing schemes from macrocells as shown in [3].
2) The non-cooperative learning process of the power control game without any information exchange case: each user's transmission decisions in the learning process are self-incentive with myopic best response correspondence, which is the scenario discussed in [10].

The first observation from our simulation results is that, whenever we generate random initial probability distributions
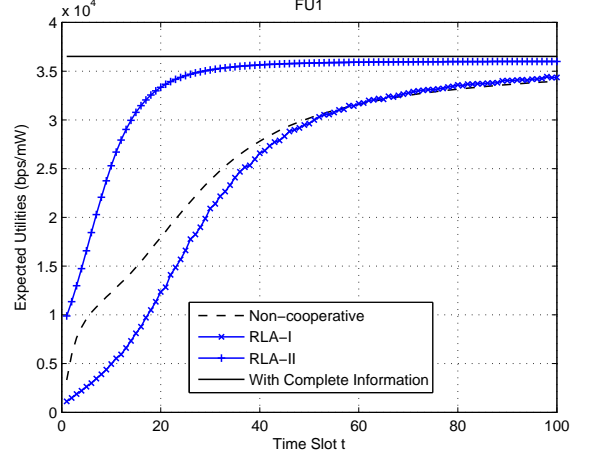

Fig. 2. Learning process of the expected utilities for FU 1.
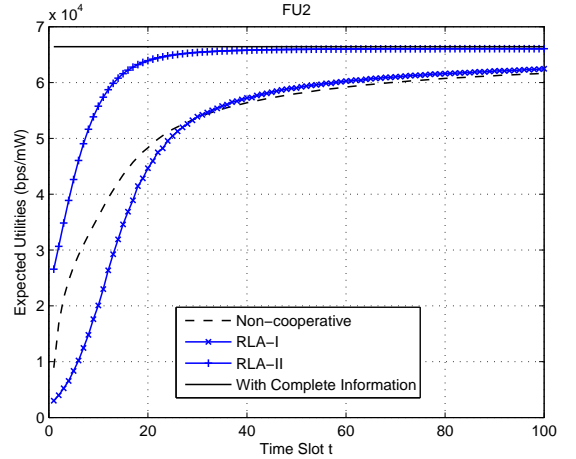

Fig. 3. Learning process of the expected utilities for FU 2.

of the power levels, the equilibrium state of the transmission strategies achieved by all users is independent of these initial values. That is, there exists an SE in the Stackelberg learning game, which confirms Theorem 2.

Secondly, we can find from the curves that the expected utilities of all users in the learning process will finally converge (or approach) to the equilibrium point in the complete cooperation case, and these simulation results validate the conclusion of Theorem 3. In addition, both the proposed reinforcement learning schemes outperform the non-cooperative case, this is because for a Stackelberg learning game, knowing more can improve not only the leader's (MU) own utility, but also the utilities of the followers (FUs). Meanwhile, the *RLA-II* can
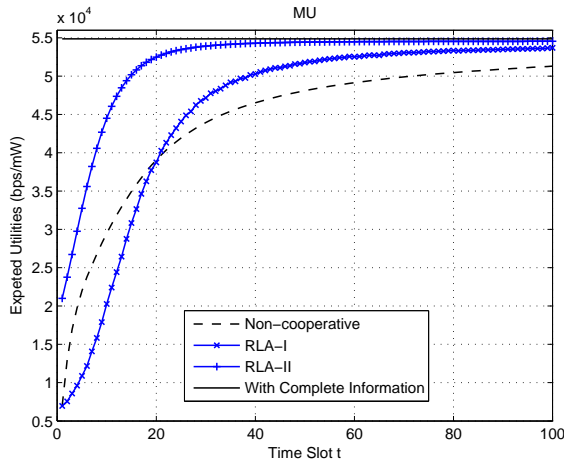
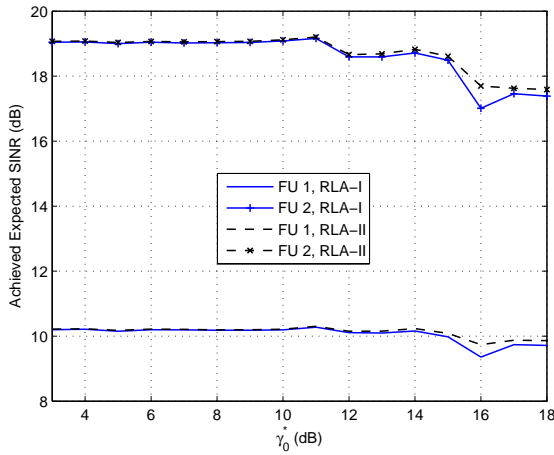Fig. 4. Learning process of the expected utilities for MU.

Fig. 5. The expected SINRs for FUs versus γ_0^*.

achieve better performance than *RLA-I*, which is due to the fact that all FUs have the incentive to achieve better utilities thus behave reciprocally at some extent.

Fig. 5 shows the expected SINRs of FUs using *RLA-I* and *RLA-II*, respectively, versus the minimum macrocell QoS requirement γ_0^*. It is observed that for the same γ_0^*, the expected SINRs of the FUs with *RLA-II* is in general higher than that with *RLA-I*. This is in accordance with our previous discussions. It is also worth mentioning that when γ_0^* is sufficiently large, the expected SINRs of the FUs approach to zero for the two learning algorithms. This is because when γ_0^* is sufficiently large, there is no femtocell active in the network.

VI. CONCLUSION

In this paper, energy efficiency is investigated for the uplink transmission in a spectrum-sharing-based two-tier femtocell network using stochastic learning theory. The Stackelberg learning framework is adopted to jointly study the utility maximization of the MU and FUs. Based on reinforcement learning, we propose two algorithms, namely, *RLA-I* and *RLA-II*, whose convergence has been further proven theoretically. Numerical experiments illustrate that the reciprocity-inspired *RLA-II* converges more quickly and achieves better performance compared to *RLA-I* and the non-cooperative learning scheme. This comes at the expense of more side information at the FUs. Concludingly, both learning algorithms show the potential in improving the energy efficiency in the femtocell networks.

REFERENCES

[1] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 30-38, October 2011.

[2] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59-67, September 2008.

[3] V. Chandrasekhar and J. G. Andrews, "Power control in two-tier femtocell networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4316-4328, August 2009.

[4] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A Stackelberg game approach," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 538-549, April 2012.

[5] R. Xie, F. R. Yu, and H. Ji, "Energy-efficient spectrum sharing and power allocation in cognitive radio femtocell networks," in *Proc. of the IEEE International Conference on Computer Communications (INFOCOM)*, Orlando, Florida USA, March 2012.

[6] I. Ashraf, L. T. W. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, Sydney, Australia, April 2010.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[8] L. Giupponi, A. M. Galindo-Serrano, and M. Dohle, "From cognition to docition: The teaching radio paradigm for distributed & autonomous deployments," *Computer Communications*, vol. 33, no. 17, pp. 2015-2020, November 2010.

[9] M. Bennis, S. Guruacharya, and D. Niyato, "Distributed learning strategies for interference mitigation in femtocell networks," in *Proc. of the IEEE Global Communications Conference (GLOBECOM)*, Houston, Texas, USA, December 2011.

[10] C. long, Q. Zhang, B. Li, H. Yang, and X. Guan, "Non-cooperative power control for wireless ad hoc networks with repeated games," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 6, pp. 1101-1112, August 2007.

[11] Y. Vorobeychik and S. Singh, "Computing stackelberg equilibria in discounted stochastic games," in *Proc. of the Twenty-Sixth National Conference on Artificial Intelligence*, Toronto, Canada, July 2012.

[12] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1992.

[13] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279-292, 1992.

[14] A. Kianercy and A. Galstyan, "Dynamics of Boltzmann Q-learning in two-player two-action games," *Physical Review E*, vol. 85, no. 4, pp. 041145, April 2012.

[15] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24 , no. 5, pp. 769-777, May 1994.

[16] A. Jean-Marie and M. Tidball, "Adapting behaviors through a learning process," *Journal of Economic Behavior & Organization*, vol. 60, no. 3, pp. 399-422, July 2006.